HUMAN-MACHINE COMMUNICATION

Human-Machine Communication
Volume 2, 2021
https://doi.org/10.30658/hmc.2.11

Out With the Humans, in With the Machines?: Investigating the Behavioral and Psychological Effects of Replacing Human Advisors With a Machine

Andrew Prahl[1] **and Lyn M. Van Swol**[2]

1 Wee Kim Wee School of Communication & Information, Nanyang Technological University, Singapore
2 Department of Communication Arts, University of Wisconsin-Madison, Madison, WI, USA

Abstract

This study investigates the effects of task demonstrability and replacing a human advisor with a machine advisor. Outcome measures include advice-utilization (trust), the perception of advisors, and decision-maker emotions. Participants were randomly assigned to make a series of forecasts dealing with either humanitarian planning (low demonstrability) or management (high demonstrability). Participants received advice from either a machine advisor only, a human advisor only, or their advisor was replaced with the other type of advisor (human/machine) midway through the experiment. Decision-makers rated human advisors as more expert, more useful, and more similar. Perception effects were strongest when a human advisor was replaced by a machine. Decision-makers also experienced more negative emotions, lower reciprocity, and faulted their advisor more for mistakes when a human was replaced by a machine.

Keywords: human-machine communication, interpersonal communication, advice, task demonstrability, emotion

Introduction

On August 13, 2014, a video titled *Humans Need Not Apply* was uploaded to YouTube and exploded in popularity, gathering over 1 million views within 3 days; it currently stands at

CONTACT Andrew Prahl • Wee Kim Wee School of Communication & Information • Nanyang Technological University • 31 Nanyang Link • Singapore 637718 • andrew.prahl@ntu.edu.sg

ISSN 2638-602X (print)/ISSN 2638-6038 (online)
www.hmcjournal.com

over 10 million views (CGPGrey, 2014; Pagano, 2014). Detailing a future in which human labor is irrecoverably taken over by automation, the 15-minute video on the evolution of labor—and why the robot revolution is different—was described as "terrifying" and foretelling "an economic horror movie" by some commentators (Roggeveen, 2014, p. 1). Since then, public interest in the continued automation of human labor has only increased; it is becoming difficult to read the business section of a popular newspaper and *not* encounter an article discussing the future of work and automation. Emerging alongside the increased discussion of machines versus humans, human-machine communication is a quickly growing subfield of communication which studies machines as interlocutors rather than simply as communication tools (Banks & de Graaf, 2020; Fortunati & Edwards, 2020). This study draws upon both interpersonal and human-machine trust literature to investigate a common real-world use for new automation: acting as an advisor to a human decision-maker. Industries as diverse as health care (Langlotz et al., 2019), finance (Lourenço et al., 2020), supply-chain management (Fildes & Goodwin, 2020), and agriculture (Zhai et al., 2020) are increasingly turning toward machines as advisors. In financial advising, for example, robo-advisors currently manage an estimated $1 trillion in assets, a number that is expected to increase to over $15 trillion by 2025 (Abraham et al., 2019; Deloitte, 2016).

The interpersonal process of trust has attracted considerable interest from scholars in a wide variety of fields (for review, see Bonaccio & Dalal, 2006). More recently, the study of human-machine trust has also increased (Hoff & Bashir, 2015; Lutz & Tamò-Larrieux, 2020). There is also scholarship emerging that attempts to explain the differences in how human and machine communicators are conceptualized (Guzman, 2020). However, little research has experimentally compared human-machine trust directly to interpersonal trust, especially in situations where machines replace humans or for decisions with more subjective and less demonstrable consequences. Given that the thought of machines replacing humans concerns many people, it is surprising how little research exists on the psychological state of people who witness machines replacing humans. The purpose of this study is to investigate (1) the effects of task demonstrability on trust in humans and machines, (2) how perceptions of advisors are affected by task type, advisor type, and advisor replacement. We begin by conceptualizing differences in task demonstrability and advisor expertise.

Task Demonstrability and Advisor Expertise

A continuum of decision-making task demonstrability anchored by intellective (high demonstrability) and judgmental tasks (low demonstrability) was explicated by Laughlin and Ellis (1986). This theoretical distinction has become important in advice research (Bonaccio & Dalal, 2006). Demonstrable tasks are distinguished by having an answer that all parties can understand. For example, an algebra problem has a correct answer and any advice provided to a decision-maker suggesting a correct answer is demonstrably correct or incorrect. On the other hand, low demonstrability tasks involve uncertain future states or subjective consequences, and the decision is seen more as a value judgment than a correct answer. Several interpersonal advice studies have varied decision-making tasks on the demonstrability continuum (Tzioti et al., 2014; Van Swol, 2011).

One of the strongest and most robust effects in advice research is that perceptions of advisor expertise directly affect trust in the advisor (Sniezek & Van Swol, 2001). Given

that task demonstrability and advisor expertise are key constructs in interpersonal advice research, human-machine communication provides an ideal context to study how the two factors may interact because the expertise of an advisor is tied to the domain and decision-making context. This suggests that perceptions of expertise for low or high demonstrability of tasks may differ for humans and machines. In the research presented here, we manipulate both advisor type (human/machine) and demonstrability of the decision-making task. Decades of research on technology acceptance, human-automation (human-machine) trust, and the growing field of machine ethics provides some theoretical insight as to how perceptions of machine attributes differ from humans.

Perceptions of Machines and Perceptions of Humans

The comparison of interpersonal and human-machine trust in tasks of varying demonstrability, especially those involving moral decisions, introduces an interesting question about the match between advisor characteristics and the context of a decision. The majority of human-machine advice research has used highly demonstrable tasks as experimental stimuli, such as what the next number is in a mathematical sequence (de Visser et al., 2016) or a yes/no question regarding the presence of military equipment in an aerial surveillance photo (Rice & Geels, 2010). There exists little research to guide our assumptions about trust of machine advisors on less demonstrable tasks. However, machines may need to make such decisions in the future, such as a self-driving car that must decide to protect vehicle occupants at the cost of endangering pedestrians (Awad et al., 2018). Although not focused on specific decisions, some scholars have investigated perceptions of various forms of machines (e.g., robots) for varying roles in society (Katz & Halpern, 2014; Takayama et al., 2008). Machines are often perceived as being more suitable for roles that do not require emotion or sensitive communication, and more suitable for roles that require memorization and unselfish service-orientation (Takayama et al., 2008).

Several perspectives in human-machine communication literature highlight the importance of the different expectations between humans and machines (Gambino et al., 2020; Guzman, 2020; Madhavan & Wiegmann, 2007). For example, according to the perfect automation schema model (Madhavan & Wiegmann, 2007), automation is expected to be high performing but invariant, whereas humans (if not as high performing) are adaptable and are expected to learn from their mistakes. Additionally, mistakes are less expected from machines in general because decision-makers do not see machines as susceptible to biases and emotions that plague human judgment (Merritt et al., 2015). On the other hand, people know that other people are not perfect. Several authors in human-robot trust have investigated questions of emotion, but there is only limited evidence that humans consider machines to possess emotion (Guzman, 2020; Kahn et al., 2012).

If human and machine advisors are perceived as having differing fundamental attributes, it will affect the perception of either advisor's capabilities. It follows that when an advisor is not assessed to have expertise (i.e., capabilities required for a certain task), the assessments of the advisor that are dependent on the task will be affected as well. Because less demonstrable tasks require value judgments that, in turn, are tied to emotion and subjective evaluation (Horberg et al., 2011), we believe that the underlying assumption that machines lack emotion will lead to lower assessment of advisor expertise in less demonstrable tasks.

Hypothesis 1a: Machine advisors will be perceived as having less expertise in less demonstrable decision tasks than more demonstrable tasks.

Perceptions of advisor expertise also affect perceptions of the advice itself. If advice comes from non-expert sources, it is perceived as less useful and less appropriate. This effect is both logical and established in past advice research (Bonaccio & Dalal, 2006).

Hypothesis 1b/c: Advice from machine advisors will be perceived as being less (1b) useful (1c) appropriate in less demonstrable than more demonstrable decision tasks.

Task, Advisor, Trust, and Perception

In the experiment below, we attempt to control factors like task that moderate the relationship between perceptions of the advice and advisor expertise, therefore we expect to see the effects of hypotheses 1a–1c reflected in our behavioral trust measure (advice-utilization) as well, especially as advice utilization and perception of advice are measured closely in time and previous research has found a strong relationship between perception of advice and use of advice (Bonaccio & Van Swol, 2014).

Hypothesis 2: Machine advice will be utilized less than human advice, in general, when the decision is less demonstrable.

Our third hypothesis is reasoned from hypotheses 1 and 2. Because different tasks may result in different advisor characteristics becoming more salient, we predict that recipient perceptions of thought process and value similarity with human advisors will be higher in less demonstrable decision tasks because the task is thought to be better suited to a decision-maker that possesses emotion.

Hypothesis 3a/b: Human/machine advisors will be perceived as having more (3a) thought process and (3b) value similarity in less demonstrable decision tasks.

Advisor Replacement

In the research study below, we manipulate advisor type, task type, and advisor replacement. There is essentially no interpersonal or machine trust research that has been conducted specifically to test the effects of advisor replacement, but because perception of humans and machine relies on different underlying assumptions, perceptual effects related to the comparison of two stimuli may be applicable to guide our expectations. When one advisor is replaced by another type, this may elicit a comparison of the two advisors that makes the perceived attributes of both more salient. Such an effect would fit into existing literature on contrast effects in communication and impression formation research (Palmer & Gore, 2014).

Contrast effects describe the process by which exposure to one target of evaluation can change the evaluation of targets presented subsequently. For example, unattractive faces are

rated as more unattractive if the evaluator is shown an attractive face before the unattractive one (Wedell et al., 1987). We are not aware of any literature to suggest that contrast effects will not extend to the evaluation of one advisor after replacing another. If a contrast effect is found, we expect it to result in our hypothesized effects of advisor type becoming stronger. For example, if a decision-maker is presented with a new machine advisor after gaining experience with a human advisor, it may result in an even stronger perception of invariance and exaggerate expectations of high performance. To be clear, we only manipulate advisor replacement, we do not replace one task with another; our hypothesis below therefore only covers effects driven by advisor perception. For brevity, we summarize these effects in the below hypotheses:

> **Hypothesis 4a/b:** Machine/human advisors will be evaluated as less expert, useful, appropriate, and similar in less/more demonstrable decision tasks when they replace human/machine advisors than when replacing another machine/human advisor.

> **Hypothesis 4c/d:** In more/less demonstrable tasks, when a machine/human advisor replaces a human/machine advisor, the machine/human advice will be utilized more than when a machine/human advisor replaces another machine/human advisor.

Decision-Maker Emotions

Our earlier discussion of emotions primarily discussed a decision-maker's perception that an advisor possesses emotions or at least the capability to understand emotions. But perceptual processes themselves are affected by emotions, and interpersonal advice research has shown decision-maker emotions to have substantial effects on trust (MacGeorge et al., 2013). In interpersonal advice research using demonstrable tasks, researchers have manipulated decision-maker emotions, finding that the induction of other-directed negative emotions (i.e., anger, frustration) resulted in less advice utilization, while other-directed positive emotions (i.e., happiness, gratitude) resulted in more utilization (Gino & Schweitzer, 2008). Such effects were also found in research using less demonstrable tasks (de Hooge et al., 2014).

Research on the effects of decision-maker emotions and trust in machines is less conclusive about the effects of emotions on trust. This is largely because advisor anthropomorphism can have strong effects on emotion (de Visser et al., 2016; Waytz et al., 2014), and there are large differences in anthropomorphism for machines (i.e., a social robot versus a calculator). Because we do not manipulate decision-maker emotion in our study, we are only able to predict potential effects that result from our manipulations advisor type and task, but our study design is ideally suited to investigate the emotions that may be produced by interacting with human versus machine advisors and the effects of replacing one advisor with another. A human being replaced by a machine, for example, could produce a negative emotional reaction due to the belief that the machine is not suited for the decision task or vice versa. Thus, we incorporate measures of positive emotions (e.g., happiness) and negative emotions (e.g., anger).

In addition to emotions, we also measure two processes related to trust: reciprocity and fault. Reciprocity—the belief of owing something to one's advisor—is interesting because trust is often conceptualized as a reciprocal process (Mayer et al., 1995). We also examine attributions of fault for mistakes because if humans are expected to be fallible and imperfect (Madhavan & Wiegmann, 2007), it may result in decision-makers generally finding less fault in human advisors' mistakes. We are also interested in fault because it is possible that decision-makers will fault machines to a greater degree than human advisors because fault may be related to *blame*. Our low demonstrability decision task in this experiment has consequences that result in the loss of human life, and though the measurement of what is perceived as "moral" is complicated, it is not unreasonable to assume that decision-makers could sense moral implications. Some research and emerging machine ethics research suggests many humans have a discomfort with placing blame on machines for making decisions with moral implications because many people do not perceive machines to possess moral accountability (Kahn et al., 2012), our experimental manipulations offer a unique opportunity to investigate this question.

RQ1/2/3: Are decision-maker (1) emotions, (2) reciprocity, and (3) attributions of fault affected by task and advisor type?

Method

Participants

Participants were recruited through Amazon's Mechanical Turk (MTurk) service and were required to be U.S. citizens over 18. Typical MTurk samples have limitations; for example, they tend to be younger and more likely to vote Democratic (Levay et al., 2016), and the use of scripts or bots is possible. To minimize potential problems, we first specified that subjects were "Master" workers (have a history of providing high quality work), and we used MTurk worker qualifications (i.e., age, gender, geographic location) to ensure a sample similar to the U.S. general population. Finally, our screening questions on the survey itself were set with quotas of demographics such as age and gender as a second layer of verification. Throughout the survey, attention and bot check questions were presented at random intervals; any subject failing two or more check questions was eliminated. Power analyses were conducted based off of past research (see appendix) and given the very slight manipulations present in our research, we recruited a large enough sample to detect effects. A total of 689 participants completed the study. In the high demonstrability task: $n = 321$, there were $n = 80$ participants in the machine advisor replaced by machine (MrM), $n = 77$ in human advisor replaced by human (HrH), $n = 84$ in machine advisor replaced by human advisor (MrH), $n = 80$ in human advisor replaced by machine advisor (HrM). Low demonstrability task: $n = 368$, $n = 82$ (MrM), $n = 74$ (HrH), $n = 104$ (MrH), $n = 108$ (HrM).

Task

A forecasting task was chosen to maximize relatability to previous research comparing human and machine advice (e.g., Fildes et al., 2006; Önkal et al., 2009), because it allowed for the clean manipulations of task demonstrability, and because forecasting is a task for

which machines are being increasingly used in the real world, for example, in supply chain forecasting (Fildes & Goodwin, 2020). Participants completed 20 forecasting tasks; all 20 graphs/forecast scenarios were randomly assigned within each task condition. In the high task demonstrability condition, the forecasting scenarios related to hospital operating room management (screenshots in appendix). In the low task demonstrability condition, the same graphs were displayed, but the scenarios dealt with humanitarian relief.[1]

Procedure

The manipulation of advisor type was simple and similar to past studies (Önkal et al., 2009; Prahl & Van Swol, 2017). Participants were told at the opening that the advice would come either from an algorithmic software program (OptiLytics), or an experienced surgeon at the hospital in the high demonstrability condition (in the low demonstrability condition, a humanitarian relief professional). The advisors were introduced to participants with a short photograph describing their/its role in the organization (see appendix for descriptions). Midway (after trial 10), the advisor was replaced with either the same type (human/machine) or different type of advisor to create the HrH, HrM, MrM, and MrH conditions.[2]

Compensation for participants was set at 25% above the federal (USA) minimum wage rate assuming a 45-minute completion time. Participants were presented graphical representations of past data, similar to a stock price chart, and then asked to make an initial forecast of where the value would be in the future. After making an "initial" forecast, the advice from a human (or algorithmic) advisor was presented; participants could make a "revised" forecast on the screen and submit it as their final forecast.[3]

At the end of each trial, a performance feedback screen was shown that displayed the participant's final forecast, the advisor's forecast, and the actual correct answer. Additionally, percentage errors were calculated for each forecast (allowing them to compare their own performance versus the advisor's performance). The participant was also shown their average percentage error across all trials so they could see if their performance on the individual trial was better or worse than previously. Finally, in the high demonstrability condition, the participant's percentage error was multiplied by 1000 and presented as (for a 1% error): "This forecasting error is estimated to have cost the hospital $1000." In the low demonstrability condition, the percentage error was multiplied by 100 and displayed as: "This forecasting error is estimated to have resulted in 100 deaths." This was to reinforce that the decisions had either financial consequences or consequences resulting in adversity to humans. The first 10 trials were performed to set the stage for the second group of 10 trials where our research interest in advisor replacement lies.

1. We picked these domains partially to control for participants having personal intuition for what the outcomes would be. Recipients often use advice less if they believe themselves to possess unique domain expertise (Lawrence et al., 2006); our task minimizes this risk.

2. When advisors were replaced midway through the survey, the introduction text was preceded by "Due to time constraints, we were not able to get [advisor]'s advice for every forecast. Therefore, you will have a new advisor to help you on the remaining tasks. Your new advisor is . . . "

3. Javascript coding was written into the survey to control the accuracy of the participant forecast and advice forecast, which was varied slightly between trials. This resulted in the relative performance of the participant and the advisor always being the same across all participants to control for related confounds.

Measures

A measure used in past research to assess behavioral trust, the "SHIFT" variable, was used to measure advice-utilization. This measure not only provides commonality with forecasting research and algorithmic advice research (e.g., Önkal et al., 2009), but also commonality with interpersonal advice studies which have used the equivalent "Weight of Advice" measure to assess trust as a behavioral measure (for review, see Bonaccio & Dalal, 2006). The SHIFT formula is:

> (Judge revised forecast – Judge initial forecast) / (Advisor forecast – Judge initial forecast)

A questionnaire was administered to assess perceptions of the advisor (e.g., expertise), advice (e.g., appropriateness), and decision-maker emotions, reciprocity, and fault after the first set of 10 trials and after the final 10 trials with the replacement advisor.[4] The survey, consisting of sematic differential and Likert style survey questions, was constructed with survey items from previous advice literature (MacGeorge et al., 2013), details of survey measures and reliability can be found in the appendix.

Results

Manipulation Checks

To confirm participants perceived the tasks differently, we conducted independent sample *t* tests on our manipulation check questions between tasks. Means and standard deviations are displayed in Table 1. The task manipulation was successful.

Hypotheses

Hypothesis 1 stated that machine advice would be perceived as (1a) less expert, (1b) less useful, and (1c) less appropriate than human advice in the humanitarian than the management task. We conducted two-way ANOVAs for perceptions of advisor expertise, appropriateness of advice, and usefulness of advice in the first block (pre-replacement) with task type and advisor type as independent variables. There was no significant interaction between task type and advisor type for expertise of advisor $F(1,669) = 0.062$, $p = 0.803$, $d = 0.062$, appropriateness of advice, $F(1,658) = 0.025$, $p = 0.873$, $d = 0.058$, or usefulness of advice, $F(1,658) = 0.003$, $p = 0.998$, $d = 0.001$. Follow-up univariate tests indicated a main effect of advisor only: human advice was always perceived as being more expert $F(1,669) = 20.681$, $p < 0.001$, $d = 0.356$, more appropriate $F(1,658) = 53.803$, $p < 0.001$, $d = 0.570$, and more useful $F(1,658) = 62.350$, $p < 0.001$, $d = 0.616$, in the first block of trials. Because the main effect of advisor type was significant in the survey results from the first block of trials,

4. The questionnaire measures were identical and measured perceptions of advice usefulness and advisor quality on a sematic differential scale. Additionally, Likert survey questions measured emotions when receiving advice, trust of advisor, similarity (value, social norm, and thought process) to advisor, and perceptions of advisor effort.

TABLE 1 Manipulation Checks Independent *t* Tests

Statement	Humanitarian Task	Management Task	Mean Difference	*p*	*d***
Task is more about Human Life	*M* = 5.573 *SD* = 1.381	*M* = 2.938 *SD* = 1.554	2.634	0.001*	1.845
Task is more about Money	*M* = 2.781 *SD* = 1.283	*M* = 4.008 *SD* = 2.190	1.226	0.001*	1.064
Task has no right answers	*M* = 2.641 *SD* = 1.269	*M* = 2.349 *SD* = 1.284	0.293	0.004	0.251
Task requires compassion	*M* = 3.424 *SD* = 1.761	*M* = 3.058 *SD* = 1.363	0.365	0.002*	0.334
Task relevant to all humans	*M* = 3.831 *SD* = 2.061	*M* = 3.265 *SD* = 1.446	0.567	0.001*	0.301
Task relevant to me personally	*M* = 2.961 *SD* = 1.060	*M* = 2.954 *SD* = 1.155	0.007	0.960	0.005

* Signifies Levine's test for equality of variances violated, equal variances not assumed test statistics used.
** Cohen's *d* effect size (maximum likelihood estimator).

we could not treat every participant as coming from the same baseline condition when completing the second survey. Thus, we created a difference score by subtracting scores in the first advisor evaluation from the second. We then conducted a 2 × 2 × 2 ANOVA with task type, advisor type, and replacement type as factors. Results indicated no significant interaction between the three factors and advice appropriateness, $F(1,659) = 1.297$, $p = 0.255$, $d = 0.086$; but there was a significant interaction for advice usefulness, $F(1,662) = 5.829$, $p = 0.016$, $d = 0.189$, and advisor expertise $F(1,662) = 3.940$, $p = 0.048$, $d = 0.155$.

To understand these interactions, we conducted two-way ANOVAs comparing the effect of replacement and advisor type on expertise and advice usefulness within each task. In the humanitarian task there was a significant interaction between replacement and advisor type for advisor expertise, $F(1,355) = 13.156$, $p < 0.001$, $d = 0.288$; this interaction was not significant in the management task condition, $F(1,305) = 0.058$, $p = 0.810$, $d = 0.005$. The interaction analyses were similar for advice usefulness: significant in the humanitarian task, $F(1,360) = 8.205$, $p = 0.004$, $d = 0.0224$, but not in the management task, $F(1,300) = 0.605$, $p = 0.437$, $d = 0.051$. An inspection of the means indicates machine advisors replacing human advisors in the humanitarian task produced the largest decrease in evaluations of advisor expertise (M = -0.439, SD = 0.714) and advice usefulness (M = -0.317, SD = 0.627), whilst human advisors replacing machine advisors produced the largest increase in ratings of expertise (M = 0.038, SD = 0.818) and advice usefulness (M = 0.063, SD = 0.766). Thus, we find partial support for hypothesis 1a and 1b, human advice is perceived as more expert and more useful than machine advice in humanitarian decision-making tasks, but only when one advisor type has replaced the other. A table detailing the three-factor ANOVAs for H1a–c, H3a–b, and our RQs can be found in Table 2; see figures for graphs of significant three-way interactions.

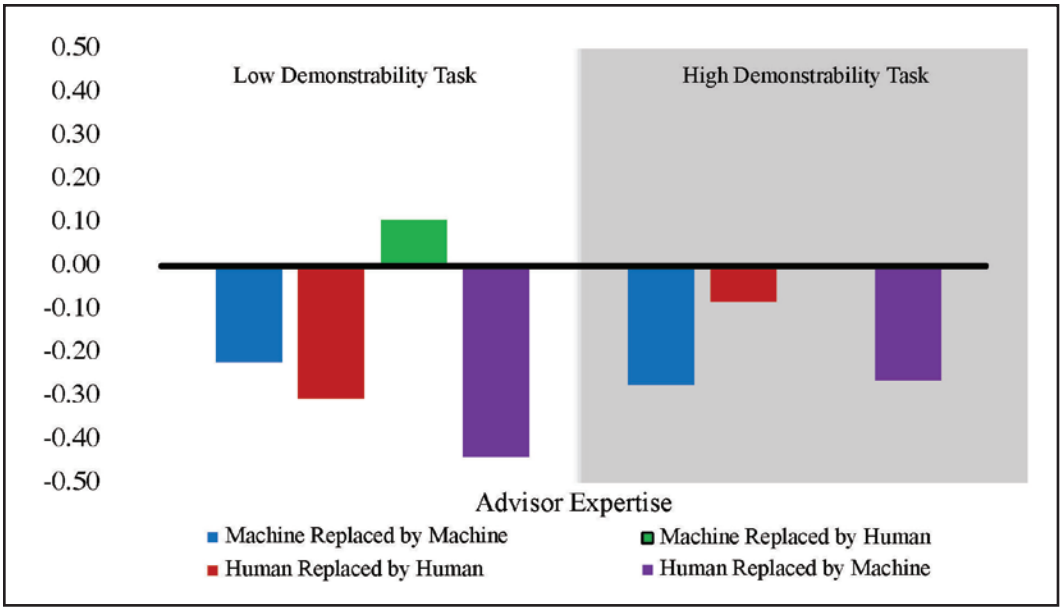


FIGURE 1 Advisor Expertise Change from Advisor 1 to Advisor 2

Positive values = more perceived advisor expertise with second advisor
Negative values = less perceived advisor expertise with second advisor

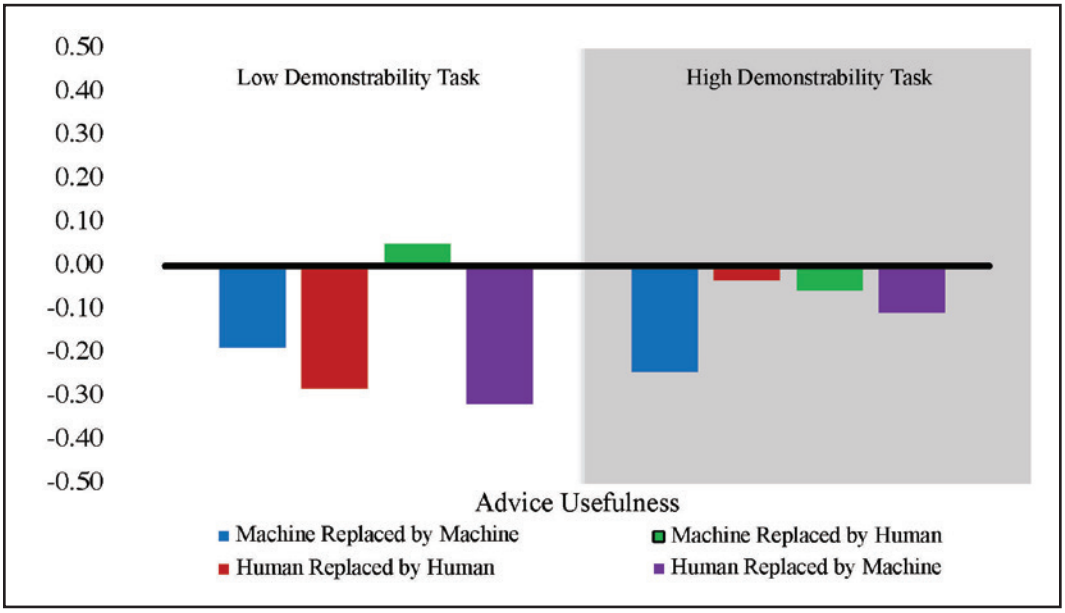


FIGURE 2 Advice Usefulness Change from Advisor 1 to Advisor 2

Positive values = more perceived advice usefulness with second advisor
Negative values = less perceived advice usefulness with second advisor

Hypothesis 2 stated that in the less demonstrable task (humanitarian), machine advice would be used less, in general, than human advice. Because all participants started the first 10 trials with a human or machine advisor, we first conducted a 2 (humanitarian/management task) × 2 (machine/human advisor) ANOVA with average advice utilization as the dependent variable for the first 10 trials. There was no significant interaction between advisor and task type $F(1,671) = 3.136$, $p = 0.077$, $\eta p^2 = 0.005$. There was a main effect of task ($F(1,671) = 8.775$, $p = 0.002$, $\eta p^2 = 0.013$) that indicated advice was used significantly more in the humanitarian ($M = 0.561$, $SD = 0.191$) than management task ($M = 0.521$, $SD = 0.188$), see Figure 1 in the appendix. To analyze the second block of trials, we had to account for similar or different advisor replacement as well as advisor type and task. We computed a variable composed of average advice utilization on the second block of trials and then conducted a 2(human/machine advisor) × 2(similar/different advisor replacement) × 2(humanitarian/management task) ANOVA but found no significant interaction between task, advisor, and replacement type, $F(1,667) = 0.656$, $p = 0.418$. We conducted a follow-up 2(humanitarian/management task) × 2(machine/human advisor) ANOVA but we did not observe a significant interaction, $F(1,667) = 0.320$, $p = 0.858$, or observe a main effect of task ($p = 0.133$) or advisor ($p = 0.602$). In sum, hypothesis 2 is not supported.

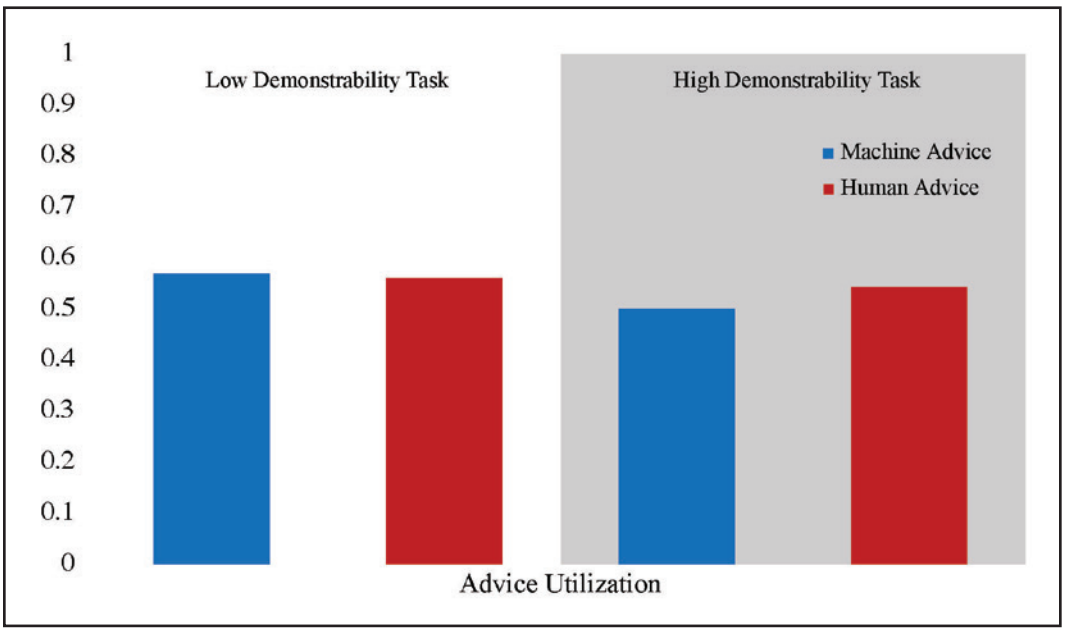


FIGURE 3 Advice Utilization First 10 Trials

Higher values = more advice utilization

Hypothesis 3a and 3b stated that perceptions of human/machine advisor thought process and value similarity would be greater/less in the humanitarian compared to the management task. For brevity, we used similar analyses to hypothesis 1, but with perceptions of value and thought process similarity. In the first block of trials there was no significant interaction between advisor and task type for perceptions of thought process similarity, $F(1,662) = 0.109$, $p = 0.741$, or value similarity, $F(1,656) = 0.256$, $p = 0.613$. Tests on the second block of trials revealed a significant three-way interaction of task, advisor, and replacement type on perceptions of thought process similarity $F(1,645) = 7.067$, $p = 0.008$, $d = 0.208$; but not on perceptions of value similarity $F(1,634) = 1.236$, $p = 0.267$. Follow-up two-way ANOVAs indicated a significant interaction between advisor and replacement type in the humanitarian task condition $F(1,352) = 38.038$, $p < 0.001$ $d = 0.672$, but in the management task condition this interaction was only marginally significant, $F(1,297) = 3.556$, $p = 0.060$, $d = 0.217$. Similar to ratings of advice usefulness (H1b), a human advisor replacing a machine advisor resulted in an increase in perceptions of thought process similarity ($M = 0.464$, $SD = 1.282$), whereas a machine advisor replacing a human resulted in the largest decrease ($M = -1.156$, $SD = 1.298$). In sum, we find partial support for hypothesis 3a, decision-makers do perceive more thought process similarity with human advisors compared to machine advisors in humanitarian decision-making scenarios, but only when one advisor type has replaced the other. We did not find support for Hypothesis 3b regarding value similarity; results are summarized in Table 2.

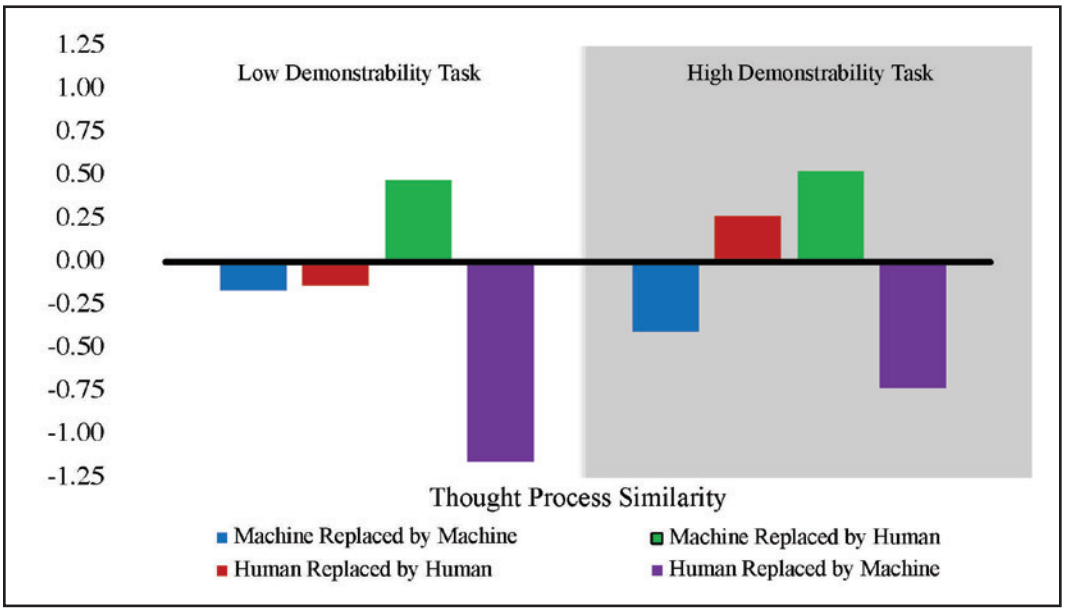


FIGURE 4 Thought Process Similarity Change from Advisor 1 to Advisor 2

Positive values = more perceived similarity with second advisor
Negative values = less perceived similarity with second advisor

TABLE 2 Difference Between Rating of First Advisor and Second Advisor, Three-Way ANOVAs and Follow-Up

Measure	Interaction	*F*	*p*	ηp^{2*}
Expertise (H1a)[1,2]	**Task*Replacement*2nd Advisor**	**3.954**	**0.047**	**0.006**
- In Humanitarian Task	**Replacement*2nd Advisor**	**13.966**	**0.001**	**0.037**
- In Management Task	Replacement*2nd Advisor	0.100	0.752	0.001
Usefulness (H1b)	**Task*Replacement*2nd Advisor**	**5.494**	**0.019**	**0.008**
- In Humanitarian Task	**Replacement*2nd Advisor**	**7.786**	**0.006**	**0.021**
- In Management Task	Replacement*2nd Advisor	0.565	0.453	0.002
Appropriateness (H1c)	Task*Replacement*2nd Advisor	1.381	0.241	0.002
Thought Process Similarity (H3a)	**Task*Replacement*2nd Advisor**	**6.817**	**0.009**	**0.011**
- In Humanitarian Task	**Replacement*2nd Advisor**	**38.369**	**0.001**	**0.099**
- In Management Task	**Replacement*2nd Advisor**	**3.931**	**0.048**	**0.013**
Value Similarity (H3b)	Task*Replacement*2nd Advisor	1.202	0.273	0.002
Positive Emotions (RQ1)	Task*Replacement*2nd Advisor	2.528	0.112	0.004
Negative Emotions (RQ1)	**Task*Replacement*2nd Advisor**	**3.888**	**0.049**	**0.006**
- In Humanitarian Task	**Replacement*2nd Advisor**	**4.114**	**0.043**	**0.011**
- In Management Task	Replacement*2nd Advisor	0.673	0.413	0.002
Reciprocity (RQ2)	Task*Replacement*2nd Advisor	0.976	0.324	0.001
Faulting the Advisor (RQ3)	**Task*Replacement*2nd Advisor**	**9.659**	**0.002**	**0.014**
- In Humanitarian Task	**Replacement*2nd Advisor**	**12.434**	**0.001**	**0.034**
- In Management Task	Replacement*2nd Advisor	0.814	0.368	0.003

[1] Significant three-way interaction graphs in Figure 2.2–5 (Appendix).
[2] **Bold** = significant at the $p = 0.05$ level.
* Effect size (partial eta-squared).

Hypothesis 4a–4d suggested that the effects posited by hypotheses above would be affected by advisor replacement, making the observed effects stronger. Our results above effectively answered these questions. We find partial support for hypothesis 4a and 4b (see results for H1a, H3a) that referred to perceptions advisor expertise, advice usefulness, advice appropriateness, and perceived advisor thought process similarity. Machine advisors

that replace human advisors are rated as having less expertise, less useful advice, and having less similar thought processes to the decision-maker when they (machine advisors) replace human advisors compared to when they replace other machine advisors in the humanitarian task, but this interaction between advisor type and replacement effect is not present in the management task. Similarly, human advisors that replace machines are rated as having more expertise, advice usefulness, and similar thought processes to the decision-maker when they (human advisors) replace machines as opposed to replacing another human; again, this effect is present in humanitarian but not management tasks. We do not find that ratings of appropriateness are significantly affected by advisor replacement. Hypotheses 4c and 4d suggested there would be an effect of advisor replacement type on utilization as well, but there was no significant interaction between task, advisor, and replacement type on advice utilization (see results for H2), hypotheses 4e and 4f are unsupported.

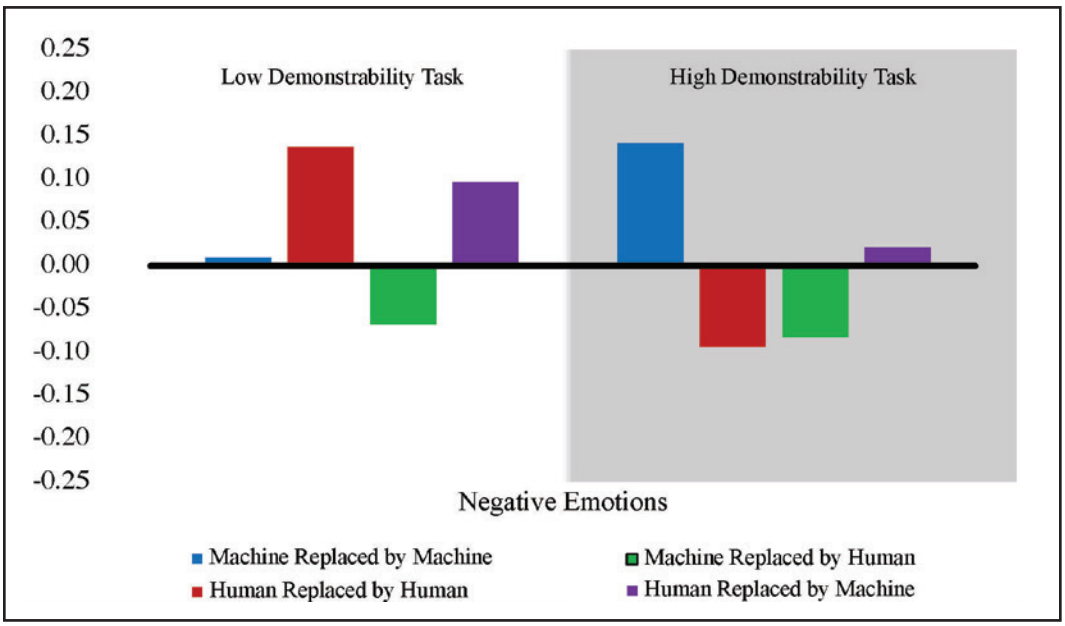


FIGURE 5 Negative Emotions Change From Advisor 1 to Advisor 2

Positive values = more negative emotions with second advisor
Negative values = less negative emotions with second advisor

With regard to RQ1 (emotions), RQ2 (reciprocity), RQ3 (fault), our analyses above suggested that the most interesting investigation would be in the difference between the rating of both advisors. We summarize the results of both exploratory paired samples *t*-tests and tests of three-way interactions between task, advisor, and replacement type in Table 3. Although these tests have not been corrected for type I error rate, we find them sufficient to answer our research questions that both decision-making task and advisor replacement type affect the emotions of decision-makers, reciprocity, and fault.

TABLE 3 Summary of paired samples *t*-tests for RQ1-3. M_{Diff} is mean difference of second advisor rating—first advisor rating (negative value = less endorsement of survey item for second advisor compared to first advisor).

Item	1	Humanitarian Task				Management Task			
		M_{Diff}	*t*	*Df*	*p*	M_{Diff}	*t*	*df*	*p*
Positive Emotion	MrM	-0.09	1.05	76	0.306	**-0.09**[2]	**2.78**	**74**	**0.001**
	MrH	0.09	1.18	102	0.246	-0.03	0.32	80	0.755
	HrM	**-0.34**	**5.13**	**107**	**0.001**	-0.16	1.56	71	0.125
	HrH	-0.15	1.41	71	0.161	-0.03	0.30	73	0.762
Negative Emotion*	MrM	0.01	0.12	78	0.90	**0.14**	**2.67**	**74**	**0.01**
	MrH	-0.05	0.78	105	0.441	-0.07	0.81	82	0.422
	HrM	**0.10**	**1.91**	**105**	**0.054**	0.02	0.27	70	0.793
	HrH	0.14	1.31	70	0.198	-0.09	1.08	74	0.292
Reciprocity	MrM	**0.26**	**2.03**	**76**	**0.052**	0.08	0.70	74	0.494
	MrH	**0.74**	**5.02**	**102**	**0.001**	**0.89**	**5.08**	**80**	**0.001**
	HrM	**-0.96**	**7.62**	**105**	**0.001**	**-0.86**	**4.22**	**71**	**0.001**
	HrH	-0.12	0.95	71	0.354	0.20	1.36	75	0.183
Fault the advisor for mistakes*	MrM	0.19	1.43	78	0.160	0.13	1.12	76	0.278
	MrH	-0.13	1.05	105	0.304	-0.08	0.58	82	0.577
	HrM	**0.70**	**4.82**	**106**	**0.001**	**0.35**	**2.16**	**72**	**0.035**
	HrH	**0.40**	**2.45**	**71**	**0.002**	-0.07	0.48	75	0.639

1: Advisor/Replacement Type: MrM = Machine Advisor replaced by Machine Advisor (similar replacement condition). MrH = Machine replaced by Human Advisor (different replacement condition), etc.
*Advisor Type × Replacement Type × Task Type three-way interaction is significant at the 0.05 level.
[2] **Bold** = significant at the $p = 0.05$ level.

Discussion

Technological innovation is leading to the increased prevalence of algorithmic, machine advice in personal and professional life for decisions of varying demonstrability in fields as diverse as medicine, financial advising, and consumer goods. Additionally, machines are increasingly replacing human workers, and this trend is being exacerbated by recent events including the Covid-19 pandemic and advancements in artificial intelligence (Hayasaki, 2020). Our results show how the field of human-machine communication can use extant research from many communication subfields to inform our understanding of an increasingly automated world. To summarize, our findings show effects relating to the perception of an advisor as well as actual advice utilization, although both sets of effects are not always related. We found support for our first hypotheses which suggested human advice would be perceived as more expert and useful than machine advice for a humanitarian relief planning decision than a management decision. However, this effect was only significant when a machine advisor had replaced a human advisor and vice versa. Our third hypothesis

suggested that advisors would also differ across tasks on decision-maker's perceptions of advisor similarity (humans perceived as more similar in humanitarian tasks). We found that human advisors were perceived as having more thought process similarity (to the decision-maker) but, again, only when the human advisor had replaced an advisor and vice versa. We did not find significant effects of task demonstrability or advisor replacement on perceptions of advice appropriateness or perceptions of advisor value similarity. There were no significant differences in advice-utilization once the decision-makers had their advisors replaced. Overall, there are not large effects of task and advisor type on utilization. Finally, our research questions showed that decision-maker emotions, reciprocity, and fault can be affected by advisor replacement and task type.

Perception and Behavior With Machines

There are a number of potential explanations for why the manipulation of advisor type, replacement, and task type may affect *perceptions* of advisors differently than affecting actual utilization *behavior*. One explanation regards the difficulty and unfamiliarity of the task. Interpersonal advice research finds that decision-makers seek and utilize advice more when they perceive tasks to be more difficult (Bonaccio & Dalal, 2006). If decision-makers in our experiment perceived the actual act of forecasting as something they were not able to do well, it may have driven the utilization of advice regardless of perception of the advisor or perceived quality of the advice. In other words, although one advisor was perceived more positively, participants may have still perceived either advisor as more informed than themselves. An interesting manipulation for future study is to select easier tasks or tasks on which decision-makers perceive themselves to have expertise. People who believe they are experts are more prone to overconfidence and advice-discounting in general (Bonaccio & Dalal, 2006), and thus, there would be a higher bar toward advice utilization.

Another potential explanation relating to expertise is the possibility that differing ratings of expertise between advisors in task conditions was due to decision-makers feeling that human advisors had more expertise than machines in understanding the consequences of a decision (human lives), but not in actually comprehending the forecasting data and producing an optimal forecast. This is an interesting direction for future study in research comparing human versus machine advice because it may uncover further underlying assumptions that decision-makers have about machine and human advisors that are specific to different parts of a decision process (Einhorn & Hogarth, 1981). In our experiment, the graph stimuli remained exactly the same, but the numbers clearly meant something different (lives versus dollars). Thus, evaluation of the numbers themselves may involve a different cognitive process (i.e., information processing) than evaluating their meaning, which is a more judgmental process. Utilization of the advice may have reflected a decision-maker's assessment of advisor's ability to perform one aspect of the decision process, but perceptions of the advisor's expertise and usefulness may reflect an assessment of a different process such as judgment. Expectations of machines play a critical role in predicting detrimental behaviors such as over- or under-reliance on machines (Madhavan & Wiegmann, 2007), and a better understanding of what aspects of the decision-making process these expectations refer to may lead to better machine advisor design and integration.

Emotion and Machines

Another potential moderator between perception and behavior is decision-maker emotions. We examined both positive and negative emotions due to past research showing emotion's effect on decision-maker perceptions and utilization (de Hooge et al., 2014). Our results showed a significant effect of task on negative emotions, but perhaps the more interesting result is that for positive emotions there was only one condition that produced an increase in positive emotions: when a human advisor replaced a machine advisor on humanitarian tasks. In the same task, machines replacing a human produced the largest decrease in positive emotions (see Table 3 for mean differences). These results should be interpreted with caution because they did not result in significant omnibus effects, but there is a clear direction implied—decision-makers are not feeling positive emotions when machines replace a human advisor when making a less demonstrable decision. In our study and interpersonal advice research in general, it is unclear how emotions are tied to utilization behavior. For example, Gino and Schweitzer (2008) found that inducing anxiety led to more advice utilization, but de Hooge et al. (2014) found that negative emotions resulted in lower perceived expertise of an advisor and lower utilization. Our results show that emotion is an important area for future research, especially because the emotional reaction to receiving advice seems to differ between human and machine advisors.

Human Similarity and Liking

Perceived advisor similarity is another set of findings that provide insight into the complicated relationship between humans and machines. Advisor thought process similarity ratings for human advisors increased more when they replaced machine advisors. Although we do not know if decision-makers are consciously comparing one advisor to the other, a large amount of contrast effect research suggests this process happens unconsciously (Palmer & Gore, 2014). Perceived similarity does generally result in more liking (Strauss et al., 2010) and the implication of this is not only that humans may like human advisors that replace machines, but that humans do not like machines that replace other humans. This is an important area for future research; there is conflicting survey evidence about how much the average people like the idea of machines replacing humans (Savela et al., 2017), and some field research suggests that machines are sometimes welcomed as a replacement to humans (Wasen, 2010) or desired not to replace humans (Kristoffersson et al., 2011).

Human advisors also may be liked more in general because our results show that decision-makers feel more reciprocity (i.e., "I owe something to my advisor") toward human advisors. This is quite a remarkable result if one considers that our manipulation of human versus machine advisor was very minimal in this experiment. While agency and influence (Banks & de Graaf, 2020) were present for the advisor, there was almost no interactivity with either advisor nor was there any conversational wording added to the advice; it was simply delivered as a number. In conjunction with our results regarding advisor perception above, this result has important implications for human-machine trust theory—especially continued efforts to investigate what degree people see machines as social actors (Gambino et al., 2020). Our results suggest that even our small manipulation with no social interaction leads to very different assessments of a social feeling like reciprocity.

Research Implications and Conclusion

Our results also have real-world implications. If a human is replaced by a machine in the workplace, the interpersonal advice process clearly is a more social process than the human advisor in this experiment. Yet, our experiment revealed this perception of a social process is substantially different for human versus machine advisors despite the advice being hardly social at all thus. In settings outside the lab, there are likely to be several differences between human and machine advice that would act as confounds if not controlled in a laboratory setting. Our experiment removed the confounds introduced by actual real-world social relationships between humans that exist in the workplace, and thus this was a very conservative comparison of humans versus machine. When humans with real relationships are replaced by machines, perhaps these elements of social interaction are not "replaced," but actually "lost" instead. Humans are social creatures and the feeling that someone is helping you is a good one. There could be serious long-term consequences to the lost positive emotions that come from social interaction, everything from organizational commitment to productivity (Oswald et al., 2015) is at risk when employees are not happy at work. Moving forward, gaining a more thorough understanding of what happens socially and emotionally when a human colleague is replaced by machines is critical.

Additional real-world implications of our study are numerous. It is clear that humans do not like it when machines replace a human advisor, even a human advisor who is zero-acquaintance and only imagined. Furthermore, our results suggest that decision-makers *really* do not like it when this replacement occurs on a task that is less demonstrable. But the negative feelings experienced when machines replace a human do not necessarily mean that the machine advisor will be used less than the human advisor. If anything, our utilization results suggested machine advisors were used more in the humanitarian task, the same task that produced the most negative evaluations of the machine advisor when it replaced a human. Understanding how the manipulation of advisor characteristics, situational context, decision-maker self-efficacy, and advice accuracy affect this complicated relationship is important if machine advisors are to be effectively introduced into the areas they are being developed for including health care, financial advising, and disaster management. These industries are just a few of the many which will see the increased presence of machine advisors—and this trend is only projected to increase as the Covid-19 pandemic has dramatically increased corporate efforts to automate workforces. In conclusion, our research shows the process of replacing human advisors with machines will be complicated. Moreover, our research shows that it is not only the humans who are replaced that will be unhappy; the people who must work with these new machines may not be happy either.

Author Biographies

Andrew Prahl (PhD, University of Wisconsin-Madison) is an Assistant Professor at the Wee Kim Wee School of Communication & Information at Nanyang Technological University, Singapore. His research addresses the communication consequences of replacing humans with machines. More broadly, Andrew's research investigates the key issues raised by the automation of labor.

https://orcid.org/0000-0003-3675-3007

Lyn M. Van Swol (PhD, University of Illinois at Urbana-Champaign) is a professor at University of Wisconsin-Madison and associate editor of Group Dynamics. Her research examines utilization of advice and acceptance of information in contexts like automation, forestry and conservation behaviors, and hospital administration. She also researches group communication and group decision-making, especially focusing on factors that increase one's influence in a group.

https://orcid.org/0000-0002-2484-748X

References

Abraham, F., Schmukler, S. L., & Tessada, J. (2019). *Robo-advisors: Investing through machines* (SSRN Scholarly Paper ID 3360125). World Bank Research and Policy Briefs No. 134881. https://papers.ssrn.com/abstract=3360125

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J., & Rahwan, I. (2018). The moral machine experiment. *Nature*, *563*(7729), 59–64. https://doi.org/10.1038/s41586-018-0637-6

Banks, J., & de Graaf, M. (2020). Toward an agent-agnostic transmission model: Synthesizing anthropocentric and technocentric paradigms in communication. *Human-Machine Communication*, *1*, 19–36. https://doi.org/10.30658/hmc.1.2

Bonaccio, S., & Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational Behavior and Human Decision Processes*, *101*(2), 127–151. https://doi.org/10.1016/j.obhdp.2006.07.001

Bonaccio, S., & Van Swol, L. (2014). Combining information and judgments. In R. S. Dalal & S. Highhouse (Eds.), *Judgment and decision making at work* (pp. 178–198). Routledge.

CGPGrey. (2014, August 13). *Humans need not apply*. https://www.youtube.com/watch?v=7Pq-S557XQU

De Hooge, I. E., Verlegh, P. W. J., & Tzioti, S. C. (2014). Emotions in advice taking: The roles of agency and valence. *Journal of Behavioral Decision Making*, *27*(3), 246–258. https://doi.org/10.1002/bdm.1801

Deloitte. (2016). *The expansion of robo-advisory in wealth management*. Deloitte Deutschland Financial Services. https://www2.deloitte.com/de/de/pages/financial-services/articles/the-expansion-of-robo-advisory-in-wealth-management.html

De Visser, E. J., Monfort, S. S., McKendrick, R., Smith, M. A. B., McKnight, P. E., Krueger, F., & Parasuraman, R. (2016). Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied*, *22*(3), 331–349. https://doi.org/10.1037/xap0000092

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General, 144(1)*, 114. https://doi.org/10.1037/xge0000033

Einhorn, H. J., & Hogarth, R. M. (1981). Behavioral decision theory: Processes of judgement and choice. *Annual Review of Psychology*, *32*(1), 53–88. https://doi.org/10.1146/annurev.ps.32.020181.000413

Fildes, R., & Goodwin, P. (2020). *Stability and innovation in the use of forecasting systems: A case study in a supply-chain company* (Department of Management Science Working Paper, Lancaster University 2020:1). https://doi.org/10.2139/ssrn.3548701

Fildes, R., Goodwin, P., & Lawrence, M. (2006). The design features of forecasting support systems and their effectiveness. *Decision Support Systems*, *42*(1), 351–361. https://doi.org/10.1016/j.dss.2005.01.003

Fortunati, L., & Edwards, A. (2020). Opening space for theoretical, methodological, and empirical issues in human-machine communication. *Human-Machine Communication*, *1,* 7–18. https://doi.org/10.30658/hmc.1.1

Gambino, A., Fox, J., & Ratan, R. (2020). Building a stronger casa: Extending the computers are social actors paradigm. *Human-Machine Communication*, *1,* 71–86. https://doi.org/10.30658/hmc.1.5

Gino, F., & Schweitzer, M. E. (2008). Blinded by anger or feeling the love: How emotions influence advice taking. *The Journal of Applied Psychology*, *93*(5), 1165–1173. https://doi.org/10.1037/0021-9010.93.5.1165

Guzman, A. (2020). Ontological boundaries between humans and computers and the implications for human-machine communication. *Human-Machine Communication*, *1,* 37–54. https://doi.org/10.30658/hmc.1.3

Hayasaki, E. (2020, June 17). Covid-19 could accelerate the robot takeover of human jobs. *MIT Technology Review*. https://www.technologyreview.com/2020/06/17/1003328/covid-19-could-accelerate-the-robot-takeover-of-human-jobs/

Hoff, K. A., & Bashir, M. (2015). Trust in automation integrating empirical evidence on factors that influence trust. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *57*(3), 407–434. https://doi.org/10.1177/0018720814547570

Horberg, E. J., Oveis, C., & Keltner, D. (2011). Emotions as moral amplifiers: An appraisal tendency approach to the influences of distinct emotions upon moral judgment. *Emotion Review*, *3*(3), 237–244. https://doi.org/10.1177/1754073911402384

Kahn, P. H., Kanda, T., Ishiguro, H., Gill, B. T., Ruckert, J. H., Shen, S., Gary, H. E., Reichert, A. L., Freier, N. G., & Severson, R. L. (2012). Do people hold a humanoid robot morally accountable for the harm it causes? *2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 33–40. https://doi.org/10.1145/2157689.2157696

Katz, J. E., & Halpern, D. (2014). Attitudes towards robots suitability for various jobs as affected robot appearance. *Behaviour & Information Technology*, *33*(9), 941–953. https://doi.org/10.1080/0144929X.2013.783115

Kristoffersson, A., Coradeschi, S., Loutfi, A., & Severinson-Eklundh, K. (2011). An exploratory study of health professionals' attitudes about robotic telepresence technology. *Journal of Technology in Human Services*, *29*(4), 263–283. https://doi.org/10.1080/15228835.2011.639509

Langlotz, C. P., Allen, B., Erickson, B. J., Kalpathy-Cramer, J., Bigelow, K., Cook, T. S., Flanders, A. E., Lungren, M. P., Mendelson, D. S., Rudie, J. D., Wang, G., & Kandarpa, K. (2019). A roadmap for foundational research on artificial intelligence in medical imaging. *Radiology*, *291*(3), 781–791. https://doi.org/10.1148/radiol.2019190613

Laughlin, P. R., & Ellis, A. L. (1986). Demonstrability and social combination processes on mathematical intellective tasks. *Journal of Experimental Social Psychology*, *22*(3), 177–189. https://doi.org/10.1016/0022-1031(86)90022-3

Lawrence, M., Goodwin, P., O'Connor, M., & Önkal, D. (2006). Judgmental forecasting: A review of progress over the last 25 years. *International Journal of Forecasting*, *22*(3), 493–518. https://doi.org/10.1016/j.ijforecast.2006.03.007

Levay, K. E., Freese, J., & Druckman, J. N. (2016). The demographic and political composition of mechanical turk samples. *SAGE Open*, *6*(1), 2158244016636433. https://doi.org/10.1177/2158244016636433

Lourenço, C. J. S., Dellaert, B. G. C., & Donkers, B. (2020). Whose algorithm says so: The relationships between type of firm, perceptions of trust and expertise, and the acceptance of financial robo-advice. *Journal of Interactive Marketing*, *49*, 107–124. https://doi.org/10.1016/j.intmar.2019.10.003

Lutz, C., & Tamò-Larrieux, A. (2020). The robot privacy paradox: Understanding how privacy concerns shape intentions to use social robots. *Human-Machine Communication*, *1*, 87–111. https://doi.org/10.30658/hmc.1.6

MacGeorge, E. L., Guntzviller, L. M., Hanasono, L. K., & Feng, B. (2013). Testing advice response theory in interactions with friends. *Communication Research*, *43*(2), 211–231. https://doi.org/10.1177/0093650213510938

Madhavan, P., & Wiegmann, D. A. (2007). Similarities and differences between human–human and human–automation trust: An integrative review. *Theoretical Issues in Ergonomics Science*, *8*(4), 277–301. https://doi.org/10.1080/14639220500337708

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management. The Academy of Management Review*, *20*(3), 709. https://doi.org/10.5465/amr.1995.9508080335

Merritt, S. M., Unnerstall, J. L., Lee, D., & Huber, K. (2015). Measuring individual differences in the perfect automation schema. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 0018720815581247. https://doi.org/10.1177/0018720815581247

Önkal, D., Goodwin, P., Thomson, M., Gönül, M., & Pollock, A. (2009). The relative influence of advice from human experts and statistical methods on forecast adjustments. *Journal of Behavioral Decision Making*, *22*(4), 390–409. https://doi.org/10.1002/bdm.637

Oswald, A. J., Proto, E., & Sgroi, D. (2015). Happiness and productivity. *Journal of Labor Economics*, *33*(4), 789–822. https://doi.org/10.1086/681096

Pagano, M. (2014, August 17). Humans need not apply: The future of jobs is robot-shaped. *The Independent*. http://www.independent.co.uk/voices/comment/humans-need-not-apply-the-future-of-jobs-is-robot-shaped-9673643.html

Palmer, J. K., & Gore, J. S. (2014). A theory of contrast effects in performance appraisal and social cognitive judgments. *Psychological Studies*, *59*(4), 323–336. https://doi.org/10.1007/s12646-014-0282-6

Prahl, A., & Van Swol, L. (2017). Understanding algorithm aversion: When is advice from automation discounted? *Journal of Forecasting*, *36*(6), 691–702. https://doi.org/10.1002/for.2464

Rice, S., & Geels, K. (2010). Using system-wide trust theory to make predictions about dependence on four diagnostic aids. *The Journal of General Psychology*, *137*(4), 362–375. https://doi.org/10.1080/00221309.2010.499397

Roggeveen, S. (2014, August 18). Humans need not apply: An economic horror movie. *The Interpreter*.http://www.lowyinterpreter.org/post/2014/08/18/Humans-need-not-apply-An-economic-horror-movie.aspx

Savela, N., Turja, T., & Oksanen, A. (2017). Social acceptance of robots in different occupational fields: A systematic literature review. *International Journal of Social Robotics*, 1–10. https://doi.org/10.1007/s12369-017-0452-5

Sniezek, J. A., & Van Swol, L. (2001). Trust, confidence, and expertise in a judge-advisor system. *Organizational Behavior and Human Decision Processes*, *84*(2), 288–307. https://doi.org/10.1006/obhd.2000.2926

Strauss, J., Barrick, M., & Connerley, M. (2010). An investigation of personality similarity effects (relational and perceived) on peer and supervisor ratings and the role of familiarity and liking. *Journal of Occupational and Organizational Psychology*, *74*(5), 637–657. https://doi.org/10.1348/096317901167569

Takayama, L., Ju, W., & Nass, C. (2008). Beyond dirty, dangerous and dull: What everyday people think robots should do. *Proceedings of the 3rd ACM/IEEE International Conference on Human Robot Interaction*, 25–32. https://doi.org/10.1145/1349822.1349827

Tzioti, S. C., Wierenga, B., & Van Osselaer, S. M. J. (2014). The effect of intuitive advice justification on advice taking. *Journal of Behavioral Decision Making*, *27*(1), 66–77. https://doi.org/10.1002/bdm.1790

Van Swol, L. (2011). Forecasting another's enjoyment versus giving the right answer: Trust, shared values, task effects, and confidence in improving the acceptance of advice. *International Journal of Forecasting*, *27*(1), 103–120. https://doi.org/10.1016/j.ijforecast.2010.03.002

Wasen, K. (2010). Replacement of highly educated surgical assistants by robot technology in working life: Paradigm shift in the service sector. *International Journal of Social Robotics*, *2*(4), 431–438. https://doi.org/10.1007/s12369-010-0062-y

Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology*, *52*, 113–117. https://doi.org/10.1016/j.jesp.2014.01.005

Wedell, D. H., Parducci, A., & Geiselman, R. E. (1987). A formal analysis of ratings of physical attractiveness: Successive contrast and simultaneous assimilation. *Journal of Experimental Social Psychology*, *23*(3), 230–249. https://doi.org/10.1016/0022-1031(87)90034-5

Zhai, Z., Martínez, J. F., Beltran, V., & Martínez, N. L. (2020). Decision support systems for agriculture 4.0: Survey and challenges. *Computers and Electronics in Agriculture*, *170*, 105256. https://doi.org/10.1016/j.compag.2020.105256

Appendix

1: Methods Information
2: Power Analysis
3: Task Screenshots

1: Additional Methods Information

Design Considerations. Our literature review revealed the need for careful design. For example, research on how the perceived machine suitability for societal roles is affected by anthropomorphism has important design implications. First, we do not anthropomorphize the machine in order to provide the cleanest manipulation of advisor type. Second, we limit the social aspects of the advice exchange process; there is no direct interaction with either advisor type, and advice is delivered in a simple text format. Additional design implications are based on above review of advisor expertise—we avoid creating implied expertise by clearly introducing the human and machine advisors as having equivalent expertise. We also precisely control the accuracy of advice to rule out the confound of an advisor actually being better at a decision-making task. Our design therefore is optimized to discover differing assumptions that people have about the attributes of machines versus humans on tasks of different demonstrability.

Survey Measures. The questionnaire measures were identical and measured perceptions of advice usefulness and advisor quality on a sematic differential scale. Additionally, Likert survey questions measured emotions when receiving advice, trust of advisor, similarity (value, social norm, and thought process) to advisor, and perceptions of advisor effort.

Positive emotions were measured with four Likert questions on a 1 (not at all) to 5 (extremely) scale for four positive emotions: Appreciative, Happy, Grateful, Thankful. The negative emotion scale was composed of Mad, Frustrated, Annoyed, Irritated. The four positive and negative emotion questions produced sufficient reliability (positive: $\alpha = 0.940$, negative: $\alpha = 0.943$), and the mean was used as an index of positive/negative emotion. Four sematic differential questions were used to measure advice usefulness (e.g., thoughtful, useful); and achieved sufficient reliability ($\alpha = 0.811$) and is hereon presented as an index of advice usefulness. Finally, in order to keep the survey a reasonable length, a pair of questions was asked to assess feelings of reciprocity to the advisor (i.e., "I feel like I owe something to my advisor for their help").

Advisor Descriptions. We pilot-tested 20 descriptions (10 human, 10 machines) with 23 undergraduate students and selected the descriptions that were closest to one another in ratings of perceived expertise, clarity, and performance expectancy.

Intellective (high demonstrability) Task: Machine. Your advisor today is a computer program called OptiLytics. OptiLytics is a software program used by the Gain Healthcare System to help with forecasting. The statistical models in OptiLytics have been built using 10 years of past Gain Healthcare data, as well as some data from the Center for Disease Control in the United States.

Intellective (high demonstrability) Task: Human. Your advisor today is Logan Girard. Logan is a medical doctor who has been working for Gain Healthcare for 10 years doing operating room management and hospital operations. Prior to joining Gain, Logan gained experience in healthcare management with the Center for Disease Control in the United States.

Judgmental (low demonstrability) Task: Machine. Your adviser today is a computer program called ReliefLytics. ReliefLytics is a computer program used by the UNHCR to help with forecasting. The statistical models in ReliefLytics have been built using 10 years of past UN data, as well as some data from the Center for Disease Control in the United States.

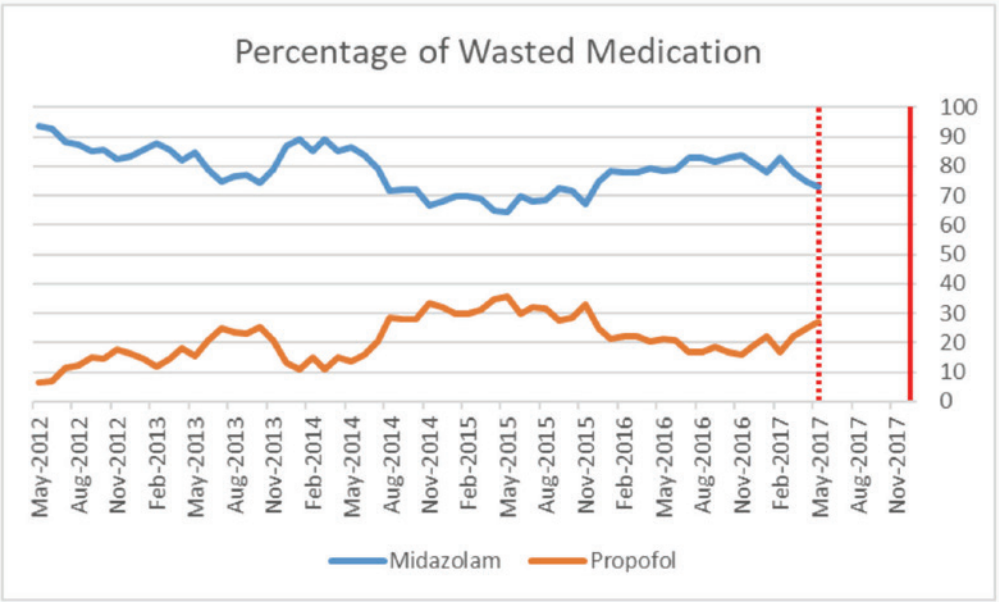
Judgmental (low demonstrability) Task: Human. Your adviser today is Logan Girard. Logan is a medical doctor who has been working for the UNHCR for 10 years doing camp management and emergency relief. Prior to joining the UNHCR, Logan gained experience managing medical crises in developing nations while working for the Center for Disease Control in the United States.

2: Power Analysis

A power analysis for the comparison between human and machine advisors was conducted using G*Power 3.1 and drew upon the three studies determined to be most similar to the research proposed (Dietvorst et al., 2015; Önkal et al., 2009; Prahl & Van Swol, 2017). Although these studies did not all report repeated measures results, the effect sizes were calculated as best as possible using published data. Dietvorst et al., reported effect sizes of 0.52 (Study 1) and 0.55 (Study 2); Önkal et al. effect size was calculated at 0.82; and Prahl & Van Swol reported a Cohen's d of 0.42. Of these, the most conservative estimate of total sample size needed by using the Prahl & Van Swol study, with G*Power calculating a needed N = 142 at alpha = 0.95 and a desired power of 0.80. Due to the well-known tendency of forecasting studies to experience high subject attrition due to missing data or the drawbacks of the weight of advice measure (for review, see Bonaccio & Dalal, 2006; Prahl & Van Swol, 2017; Tzioti et al., 2014) the target n is 162 in each advisor/task condition (human/machine & high/low demonstrability), leading to a total N = 648. Given the lack of previous studies, we have no power analyses for the advisor replacement effects, but subjects will be split into replacement conditions in each advisor/task condition and, given equivalent effect sizes, the above sample should be adequate.

3.1: Task Screenshots 1: Initial Forecast

Propofol and Midazolam are common drugs used to sedate patients, but often too much is prepared for the surgery and not used, resulting in us having to throw away some very expensive materials! Management wants to order less of sedative drugs in December 2017 to reduce waste.



Please help us plan this reduction: what percent of wasted sedative drugs that you think would be from Midazolam (blue line) in December 2017?

3.2: Task Screenshots 2: Feedback Screen